Temporal CAMML: Confidence-Aware Temporal Motion Masking for Self-Supervised Autism Behavior Representation Learning

***Abstract*—Video-based analysis of autism behavior can help with the objective assessment of repetitive and stereotypical behavior. However, existing skeleton-based approaches are mostly dependent on fully supervised learning and large annotated datasets, which are hard to obtain for autism research. To address this, we propose Temporal CAMML, a self-supervised representation learning framework for autism behavior analysis based on confidence-aware temporal motion masking. In the proposed framework, the 17-keypoint pose sequences from videos are first extracted by YOLO11-Pose. Then each behavior event is represented by a normalized temporal pose tensor. Temporal CAMML differs from traditional masking techniques that remove individual joints by identifying low-confidence skeletal observations and masking connected temporal segments of motion, thus encouraging the model to learn robust motion dynamics. A transformer encoder is then trained to reconstruct the masked trajectories, allowing for self-supervised representation learning without behavior labels. We then evaluate the learned representations with a linear probe on downstream behavior classification . Experiments on the self-stimulatory behavior data with 102 annotated events of arm flapping, head banging and spinning demonstrate the effectiveness of the proposed approach. Temporal CAMML achieves 61.9% classification accuracy, outperforming a supervised Transformer baseline (52.38%) and a confidence-aware masking variant without temporal motion modeling (52.38%). Our results demonstrate that confidence-aware temporal masking facilitates learning more discriminative pose representations for autism behavior analysis in low-resource settings.**

***Index Terms*—Autism behavior analysis, self-supervised learning, pose representation learning, temporal masking, Transformer.**

I. Introduction

Autism Spectrum Disorder (ASD) is a neuro-developmental condition characterized by differences in social communication and the presence of restricted and repetitive behaviors [16] [14]. Among these behaviors, self-stimulatory actions such as arm flapping, head banging, and spinning are frequently observed and often serve as important indicators during clinical assessment. Traditionally, the evaluation of these behaviors relies on observations made by clinicians and caregivers. While effective, such assessments can be subjective, time-intensive, and challenging to apply consistently across large populations [15]. The growing availability of computer vision and machine learning [17] techniques has enabled the automated analysis of human behavior from video data. Within this context, skeleton-based representations have emerged as an effective approach for modeling movement patterns. Unlike appearance-based methods, pose-based representations focus on the spatial and temporal relationships of body joints, making them less affected by variations in lighting conditions, background environments, and individual appearance. By representing human motion as sequences of skeletal joints, these methods provide an interpretable and efficient framework for capturing repetitive movement patterns, making them particularly suitable for the analysis of self-stimulatory behaviors associated with ASD. Despite the success of skeleton-based behavior recognition methods, most existing approaches rely heavily on supervised learning [18]. These methods require large amounts of annotated data to learn discriminative motion representations effectively. However, collecting and annotating autism-related behavioral datasets remains both costly and challenging due to privacy concerns, limited data availability, and substantial variability in behavioral manifestations across individuals. Consequently, many autism datasets are relatively small, restricting the performance and generalizability of fully supervised deep learning models [19]. To address the scarcity of labeled data, self-supervised learning (SSL) has emerged as an effective paradigm for representation learning from unlabeled samples. SSL leverages pretext tasks that encourage models to capture meaningful structural and temporal patterns within the data without requiring manual annotations . While SSL has demonstrated remarkable success across natural language processing, computer vision, and human action recognition, its application to skeleton-based analysis of autism-related behaviors remains largely underexplored. Among SSL approaches, masked representation learning has gained considerable attention, where portions of the input sequence are intentionally concealed and the model is trained to reconstruct the missing information. Existing masking strategies typically operate on randomly selected joints or isolated spatial locations [1]. However, self-stimulatory behaviors associated with ASD are inherently temporal and are characterized by repetitive and structured motion trajectories rather than individual pose configurations. Consequently, masking isolated joints may fail to capture the motion dynamics that distinguish these behaviors [20].

To overcome these limitations, we propose Temporal CAMML, a self-supervised framework designed to learn robust representations of autism-related behaviors from skeleton sequences. The proposed framework begins by using YOLO11-Pose to extract skeletal keypoints from video recordings, transforming each behavioral event into a normalized se-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

quence of poses over time. To improve representation learning, the method identifies skeletal observations with low confidence and uses this information to mask entire motion segments rather than individual joints. This temporal masking strategy encourages the Transformer-based encoder to reconstruct missing motion patterns, allowing it to learn meaningful behavioral representations without relying on manual labels. The framework is evaluated on the self-stimulatory behavior data (SSBD), which includes annotated examples of arm flapping, head banging, and spinning behaviors commonly observed in autism. Results show that Temporal CAMML produces richer and more discriminative motion representations than both a fully supervised Transformer model and a confidence-aware masking approach that masks observations without considering temporal motion segments. These gains lead to improved behavior classification performance, suggesting that confidence-aware temporal masking is an effective strategy for skeleton-based autism behavior analysis, particularly in scenarios where labeled training data are limited. The main contributions of this work are summarized as follows:

- We introduce Temporal CAMML, a confidence-aware temporal motion masking strategy for self-supervised autism behavior representation learning.
- We develop a Transformer-based self-supervised framework that learns motion representations from masked pose trajectories without requiring behavior annotations.
- We construct a skeleton-based autism behavior analysis pipeline using YOLO11-Pose extraction and normalized temporal pose representations derived from SSBD videos.
- We demonstrate through experimental evaluation that Temporal CAMML learns more discriminative representations than both supervised Transformer and non-temporal confidence-aware masking baselines, resulting in improved autism behavior classification performance.

II. Literature Survey

Artificial intelligence has increasingly been applied to the analysis and diagnosis of Autism Spectrum Disorder (ASD) using behavioral observations, neuroimaging data, and video recordings. In the context of behavioral analysis, Dia *et al.* [1] investigated deep learning approaches for continuous affect recognition in children with ASD from video sequences, demonstrating the potential of automated video-based assessment. Similarly, Jeyarani and Senthilkumar [2] proposed DSVTN-ASD, a dual self-supervised video transformer framework for recognizing stereotypical behaviors, highlighting the value of self-supervised representation learning for autism-related behavior understanding. A substantial body of research has also focused on neuro-imaging-based ASD diagnosis. Jha *et al.* [3] introduced a multi-head cross-attention framework for multi-site MRI analysis, while Abu-Doleh *et al.* [4] and Arunkumar and Ilango [5] developed deep learning models for ASD identification using structural brain imaging data. Graph-based approaches have further demonstrated promising results, with Zeraati and Davoodi's ASD-GraphNet [10] leveraging functional connectivity networks and graph learning techniques to improve diagnostic performance on ABIDE fMRI data. Recent studies have increasingly explored self-supervised and contrastive learning strategies to reduce reliance on labeled medical datasets. Wang *et al.* [6] proposed a self-supervised graph contrastive learning framework with diffusion-based augmentation for brain disorder detection, whereas Roy *et al.* [7] provided a comprehensive review of contrastive learning techniques for medical imaging applications, particularly under class-imbalanced conditions. These approaches demonstrate the effectiveness of representation learning from unlabeled data in healthcare settings. To address data scarcity, few-shot learning methods have also gained attention. Dhinagar *et al.* [11] presented a site-agnostic meta-learning framework for ASD classification from structural MRI, achieving strong generalization across previously unseen imaging sites. Likewise, the Cascading Patch-Wise Network proposed in [12] combined token selection with self-supervised learning to enhance few-shot image classification by exploiting both salient and non-annotated regions. Beyond ASD-specific studies, advances in human activity recognition have contributed valuable insights into temporal representation learning. Marhuenda *et al.* [8] proposed an optical-flow-based framework for activity recognition, while Lopez-Martinez *et al.* [9] investigated wearable eye-tracking technologies coupled with machine learning for early autism detection. Taken together, existing research highlights the effectiveness of self-supervised learning, graph-based modeling, temporal representation learning, and few-shot adaptation for ASD-related analysis. Nevertheless, current studies have primarily focused on neuro-imaging data, video-level behavior recognition, or generic representation learning frameworks. The use of confidence-aware self-supervised learning for modeling skeletal motion trajectories associated with repetitive autism behaviors remains largely unexplored. To bridge this gap, we propose Temporal CAMML, a confidence-aware temporal motion masking framework that learns discriminative behavioral representations through the self-supervised reconstruction of masked pose sequences.

III. Methodology

This section describes the proposed Temporal CAMML (Confidence-Aware Temporal Motion Masking) framework for self-supervised autism behavior representation learning. The framework is designed to learn discriminative motion representations from pose sequences without requiring behavior labels during pretraining. For each video event containing repetitive autism behaviors, human skeletal keypoints are extracted using a pose estimation model and transformed into normalized temporal pose sequences. A confidence-aware temporal masking strategy is then employed to remove selected motion segments, while a Transformer-based encoder learns to reconstruct the masked trajectories. The learned latent representations are subsequently used for downstream autism behavior classification.

Let a behavior event be represented as a sequence of T frames: $[V = I_1, I_2, \ldots, I_T,]$; where I_t denotes the video frame at time step t. For each frame, the pose detector

extracts J body joints. Each joint is represented by its spatial coordinates and confidence score: $[p_{t,j} = (x_{t,j}, y_{t,j}, c_{t,j}),]$; where $x_{t,j}$ and $y_{t,j}$ denote the pixel coordinates of joint j in frame t, and $c_{t,j} \in [0, 1]$ represents the detector confidence.

The complete pose sequence is represented as

$$P = \{p_{t,j}\}_{t=1,j=1}^{T,J} \in \mathbb{R}^{T \times J \times 3}, \tag{1}$$

where P denotes the temporal pose tensor, T is the number of frames, J is the number of body joints, and each pose observation consists of spatial coordinates and a confidence score. In this work, YOLO11-Pose is used to extract $J = 17$ body keypoints from each frame. To obtain a fixed-length representation suitable for batch processing and Transformer-based learning, each behavior event is uniformly sampled to $T = 64$ frames. Consequently, the final pose tensor is represented as

$$P \in \mathbb{R}^{64 \times 17 \times 3}, \tag{2}$$

where the three channels correspond to the normalized horizontal coordinate, normalized vertical coordinate, and pose confidence score, respectively.

A. SSBD Event Extraction and Pose Generation

The Self-Stimulatory Behavior Dataset (SSBD) comprises annotated videos of repetitive autism-related behaviors, including arm flapping, head banging, and spinning. Each video includes XML annotations that specify the temporal boundaries of behavioral events. Based on these annotations, individual behavior clips are extracted using their corresponding start and end timestamps. Let the i^{th} behavioral event shown as:

$$E_i = I_{t=s_i}^{e_i}, \tag{3}$$

where s_i and e_i denote the start and end frame indices of the event, respectively. For each extracted event clip, human skeletal keypoints are obtained using the YOLO11-Pose detector. Given a frame I_t, the detector may identify multiple individuals. To ensure consistency across the dataset, only the individual with the largest visible body area is retained. Let the bounding box associated with person k be represented as

$$B_k = (x_k, y_k, w_k, h_k), \tag{4}$$

where (x_k, y_k) denotes the top-left corner of the bounding box, while w_k and h_k represent its width and height, respectively. For each detected person, the bounding-box area is computed as

$$A_k = w_k h_k, \tag{5}$$

where A_k denotes the visible body area of person k. The individual selected for pose extraction is determined according to

$$k^* = \arg\max_k A_k, \tag{6}$$

where k^* denotes the selected subject. For the selected individual, YOLO11-Pose extracts 17 anatomical keypoints corresponding to the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles. Figure 2 presents an example of the pose extraction process using YOLO11-Pose on an SSBD video frame. The detected skeletal structure is then used to construct the temporal pose representations employed by the proposed Temporal CAMML framework. As a result, each frame is represented by a skeletal configuration consisting of 17 joints along with their associated confidence scores. Since behavior clips vary in duration, the resulting pose sequences exhibit different temporal lengths. To obtain a fixed-length representation suitable for self-supervised learning, uniform temporal sampling is applied. Given an original sequence containing L frames, a subset of T frames is selected according to

$$t_n = \left\lfloor \frac{n(L-1)}{T-1} \right\rfloor, \qquad n = 0, 1, \ldots, T-1, \tag{7}$$

where t_n denotes the sampled frame index. In this work, the target sequence length is fixed to $T = 64$ frames. The sampled pose sequence is subsequently converted into the fixed-length tensor representation defined in Eq. (2) and used as input to the proposed Temporal CAMML framework.

B. Pose Normalization and Confidence-Aware Representation

The skeletal coordinates extracted by YOLO11-Pose are represented in pixel space and are therefore influenced by the spatial resolution of the input video. Using raw coordinates directly can introduce variations arising from camera position, image dimensions, and subject location. To ensure greater consistency across video clips, all skeletal coordinates are normalized with respect to the frame dimensions. The overall architecture of the proposed Temporal CAMML framework is illustrated in Fig. 1. Let W and H denote the width and height of a video frame, respectively. For a detected joint $p_{t,j} = (x_{t,j}, y_{t,j}, c_{t,j})$, the normalized horizontal and vertical coordinates are computed as

$$\hat{x}_{t,j} = \frac{x_{t,j}}{W}, \tag{8}$$

$$\hat{y}_{t,j} = \frac{y_{t,j}}{H}, \tag{9}$$

where $\hat{x} * t, j$ and $\hat{y} * t, j$ denote the normalized spatial coordinates. Since the confidence values generated by YOLO11-Pose are already bounded within the interval $[0, 1]$, they are retained without additional normalization. Consequently, each normalized joint is represented as

$$\hat{p}_{t,j} = (\hat{x} * t, j, \hat{y} * t, j, c_{t,j}), \tag{10}$$

where $c_{t,j}$ denotes the confidence score associated with joint j at time step t. The complete normalized pose sequence is shown as

$$\hat{P} = \hat{p} * t, j * t = 1, j = 1^{T,J} \in \mathbb{R}^{T \times J \times 3}, \tag{11}$$

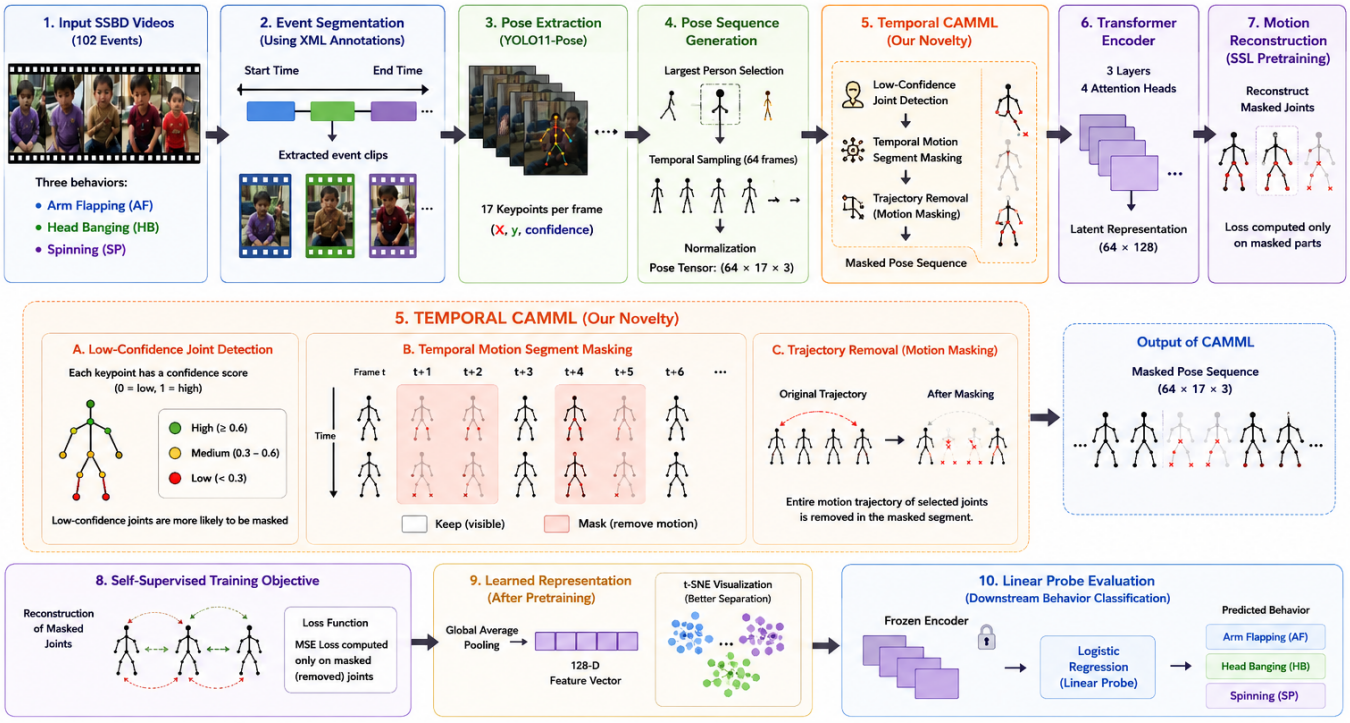


Fig. 1. Overview of the proposed Temporal CAMML framework for self-supervised autism behavior representation learning from skeletal pose sequences.

where T denotes the number of frames and J denotes the number of body joints. In this work, the normalized pose tensor is represented using the fixed-length formulation defined in Eq. (2). The inclusion of confidence values provides additional information regarding the reliability of skeletal observations. Rather than discarding these estimates, Temporal CAMML explicitly incorporates confidence information into the pose representation and subsequently utilizes it to guide the proposed confidence-aware temporal motion masking strategy described in the next subsection.

C. Confidence-Aware Temporal Motion Masking

Conventional masking strategies generally remove random joints or isolated spatial observations. However, repetitive autism behaviors are driven by structured temporal motion patterns, while pose estimation models may generate unreliable joint predictions due to occlusion, motion blur, or viewpoint variations. To improve representation learning, Temporal CAMML introduces a confidence-aware temporal motion masking strategy that focuses on low-confidence joints and masks contiguous motion segments instead of individual observations. Given a normalized pose sequence $\hat{P} \in \mathbb{R}^{T\times J\times 3}$, the confidence score associated with joint j across the entire sequence is computed by averaging the confidence values over all frames:

$$\frac{1}{T}\sum_{t=1}^{T} c_{t,j}, \tag{12}$$

where $c_{t,j}$ denotes the confidence score of joint j at frame t. The joints are subsequently ranked according to their average confidence values. The set of low-confidence joints selected for masking is defined as

$$\mathrm{TopK}_{\min}\left(\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_J\right), \tag{13}$$

where $\mathrm{TopK}_{\min}$ returns the K joints with the lowest confidence values. In this work, $K = 4$. For each selected joint $j \in \mathcal{J}_m$, a temporal segment of length L_m is randomly sampled:

$$s_j \sim \mathcal{U}\left(0, T - L_m\right), \tag{14}$$

where s_j denotes the starting frame of the masked segment and $\mathcal{U}$ represents a discrete uniform distribution. The corresponding temporal interval is given by

$$s_j, s_j + 1, \ldots, s_j + L_m - 1, \tag{15}$$

where $L_m = 8$ frames in the proposed implementation. A binary masking matrix is then constructed as

$$\begin{cases} 1, & t \in \mathcal{T}_j,\ j \in \mathcal{J}_m, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

where $M_{t,j} = 1$ indicates that the corresponding joint observation is masked. The final masked pose sequence is obtained according to

$$(1 - M) \odot \hat{P}, \tag{17}$$

where $\odot$ denotes element-wise multiplication and $\tilde{P}$ represents the masked pose tensor used for self-supervised training. By masking temporally contiguous trajectories of low-confidence joints, the proposed strategy encourages the model

to infer missing motion dynamics from both spatial context and temporal dependencies, thereby learning more robust behavioral representations.

D. Transformer-Based Motion Reconstruction Network

The masked pose sequence is transformed into a temporal feature sequence and processed using a Transformer encoder to learn motion representations. Given the masked pose tensor $\tilde{P} \in \mathbb{R}^{T\times J\times 3}$, the joint coordinates are flattened along the spatial dimension to obtain

$$X \in \mathbb{R}^{T\times 3J}, \tag{18}$$

which corresponds to $X \in \mathbb{R}^{64\times 51}$ in our implementation. The flattened pose sequence is projected into a latent embedding space using a learnable linear transformation:

$$Z_0 = XW_e + b_e, \tag{19}$$

where W_e and b_e denote learnable parameters. The embedded sequence is subsequently processed by a multi-layer Transformer encoder:

$$Z_L = \text{Transformer}(Z_0), \tag{20}$$

where Z_L denotes the learned latent motion representation.

Finally, a reconstruction layer maps the latent representation back to the original pose space:

$$\hat{X} = Z_L W_r + b_r, \tag{21}$$

where $\hat{X}$ represents the reconstructed pose sequence. By reconstructing temporally masked trajectories, the network learns motion-aware representations that capture the temporal dynamics of repetitive autism behaviors without requiring manual annotations.

E. Self-Supervised Reconstruction Objective

The proposed Temporal CAMML framework is trained using a self-supervised reconstruction objective. Given the original normalized pose sequence $\hat{P}$ and its masked counterpart $\tilde{P}$, the Transformer encoder generates a reconstructed pose sequence $\hat{X}$ that aims to recover the missing motion trajectories. To ensure that learning focuses on the masked regions, reconstruction error is computed only for the masked joint observations. Let $M \in \{0,1\}^{T\times J}$ denote the binary masking matrix defined in Eq. (**??**), where $M_{t,j} = 1$ indicates a masked joint and $M_{t,j} = 0$ otherwise. The reconstruction error for each joint observation is computed as

$$E_{t,j} = \left\| X_{t,j} - \hat{X}_{t,j} \right\|_2^2, \tag{22}$$

where $X_{t,j}$ and $\hat{X}_{t,j}$ denote the original and reconstructed joint representations, respectively. The Temporal CAMML reconstruction loss is then defined as

$$\mathcal{L}_{\text{CAMML}} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J} M_{t,j} E_{t,j}}{\sum_{t=1}^{T}\sum_{j=1}^{J} M_{t,j}}, \tag{23}$$

which corresponds to the mean squared reconstruction error over masked observations only. By restricting optimization to masked motion segments, the model is encouraged to infer missing trajectories from surrounding temporal context rather than simply reproducing visible pose information. Consequently, the learned latent representations capture meaningful behavioral dynamics that can be transferred to downstream autism behavior analysis tasks.

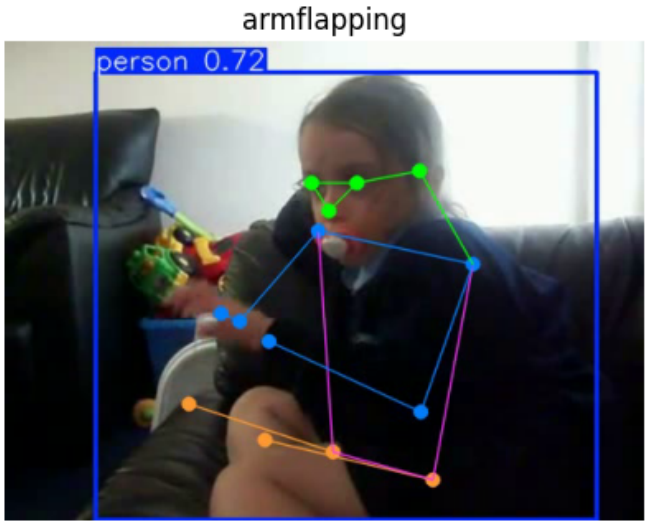


Fig. 2. Example of pose extraction using YOLO11-Pose on an SSBD behavioral event. The detected subject is represented by 17 skeletal keypoints and their corresponding confidence scores.

F. Downstream Representation Learning and Behavior Classification

After self-supervised pretraining, the Transformer encoder is used as a feature extractor for downstream autism behavior recognition. Given a normalized pose sequence, the pretrained encoder produces a latent representation

$$Z_L = \text{Transformer}(X), \tag{24}$$

where Z_L denotes the learned motion embedding. To obtain a fixed-length representation for classification, temporal average pooling is applied across all frames:

$$z = \frac{1}{T}\sum_{t=1}^{T} Z_L^{(t)}, \tag{25}$$

where $z \in \mathbb{R}^d$ represents the final behavior embedding. The quality of the learned representations is evaluated using a linear probing protocol. A linear classifier is trained on the frozen embeddings to predict autism behavior categories:

$$\hat{y} = \text{Softmax}(W_c z + b_c), \tag{26}$$

where W_c and b_c denote the classifier parameters.

In addition to linear probing, the learned representations are further evaluated using a supervised Transformer-based

behavior classification model. The classification performance provides an indirect measure of the discriminative capability of the representations learned through the proposed Temporal CAMML framework.

IV. Result Analysis and Discussion

A. Dataset Description and Implementation Details

Experiments were conducted using the Self-Stimulatory Behavior Dataset (SSBD) [13], a publicly available dataset containing videos of repetitive behaviors commonly exhibited by children with Autism Spectrum Disorder (ASD). The dataset includes three behavior categories: arm flapping, head banging, and spinning.

Using the provided XML annotations, individual behavioral events were extracted and converted into pose sequences. A total of 102 annotated events were obtained, consisting of 49 arm-flapping events, 27 spinning events, and 26 head-banging events. Each event was represented using a fixed-length pose sequence of 64 frames, where each frame contained 17 skeletal keypoints extracted using YOLO11-Pose. All experiments were implemented using PyTorch. Human pose estimation was performed using YOLO11-Pose, which extracted 17 skeletal keypoints from each video frame. To obtain a fixed-length representation, each behavioral event was uniformly sampled to 64 frames and represented as a tensor of size $64 \times 17 \times 3$. Table I summarizes the implementation settings used throughout the experiments. For self-supervised pretraining, the proposed Temporal CAMML framework selected four low-confidence joints and masked temporal segments of length eight frames. The dataset was divided into training and validation sets using an 80:20 split. All reported results were obtained on the validation set.

TABLE I
Implementation settings of the proposed Temporal CAMML framework.

Parameter	Value
Pose Detector	YOLO11-Pose
Number of Joints	17
Frames per Event	64
Embedding Dimension	128
Transformer Layers	3
Attention Heads	4
Masked Joints (K)	4
Temporal Segment Length	8
Optimizer	AdamW
Learning Rate	1×10^{-4}
Training Samples	81
Validation Samples	21

B. Evaluation Protocol and Self-Supervised Pretraining Performance

The effectiveness of the learned representations was evaluated using a linear probing protocol. Following self-supervised pretraining, the Transformer encoder was frozen and employed as a feature extractor. A linear classifier was then trained on the extracted embeddings to predict autism behavior categories. For comparison, a supervised Transformer-based classifier was trained directly on the pose sequences and evaluated against the proposed Temporal CAMML framework. Classification performance was measured using validation accuracy across three behavior classes: arm flapping, head banging, and spinning.

The proposed Temporal CAMML framework was first evaluated in a self-supervised setting through the reconstruction of masked pose trajectories. Figure 3 shows the training and validation reconstruction losses observed during pretraining. Both curves decrease steadily throughout training, indicating stable optimization and effective learning of temporal motion patterns. The reduction in validation loss suggests that the proposed confidence-aware temporal masking strategy facilitates meaningful representation learning from partially observed pose sequences. By reconstructing masked motion segments, the Transformer encoder learns temporal dependencies and behavior-specific movement dynamics without requiring manual behavior annotations. These observations indicate that the proposed self-supervised objective produces informative latent representations that can be effectively transferred to downstream autism behavior classification tasks.

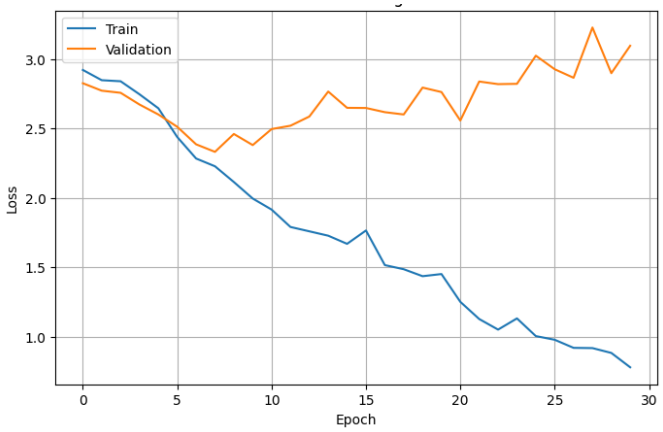


Fig. 3. Training and validation reconstruction losses during Temporal CAMML self-supervised pretraining.

C. Comparison with Baseline Methods

The proposed Temporal CAMML framework was compared against two baseline approaches: (i) a Supervised Transformer trained directly on labeled pose sequences without self-supervised pre-training, and (ii) Random CAMML, which employs random joint masking instead of the proposed confidence-aware temporal masking strategy. These baselines were selected to assess the contribution of confidence-aware masking and temporal motion reconstruction within the proposed framework. The classification results obtained by all methods are presented in Table II. Among the evaluated approaches, Temporal CAMML achieved the best validation accuracy of 61.90%, surpassing both the Supervised Transformer and Random CAMML baselines, each of which reached 52.38%. The performance gain suggests that incorporating

confidence information into the masking process helps the model learn more informative motion representations. Unlike Random CAMML, which masks joints without considering the reliability of the estimated poses, Temporal CAMML focuses on low-confidence motion trajectories and reconstructs the missing segments using information from surrounding temporal frames. This reconstruction process encourages the encoder to capture meaningful behavioral patterns and temporal dependencies, resulting in representations that are better suited for downstream autism behavior classification.

TABLE II
COMPARISON OF BEHAVIOR CLASSIFICATION ACCURACY OBTAINED USING DIFFERENT LEARNING STRATEGIES.

Method	Accuracy (%)
Supervised Transformer	52.38
Random CAMML	52.38
Temporal CAMML (Proposed)	**61.90**

Figure 4 provides a visual comparison of the obtained classification accuracies. The proposed Temporal CAMML framework consistently achieves superior performance, highlighting the benefits of confidence-aware temporal reconstruction for self-supervised autism behavior representation learning.

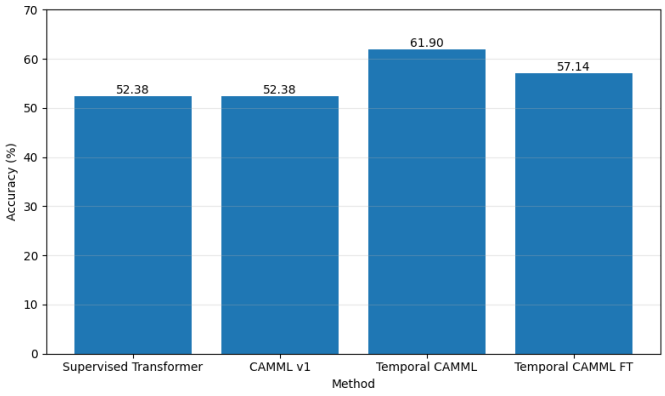


Fig. 4. Comparison of behavior classification accuracy achieved by different learning strategies.

D. Representation Analysis and Discussion

To gain further insight into the learned representations, the embeddings produced by the pretrained Temporal CAMML encoder were visualized using a two-dimensional projection technique. Figure 5 shows the distribution of behavior representations in the learned feature space. Samples belonging to the same behavioral category exhibit more compact grouping, suggesting that the self-supervised objective captures meaningful behavioral characteristics. This clustering pattern indicates that confidence-aware temporal masking encourages the encoder to learn discriminative motion dynamics rather than relying only on local pose information. The quantitative results presented in Table II are consistent with this observation. Temporal CAMML achieved a validation accuracy of 61.90%, outperforming both the Supervised Transformer and Random CAMML baselines by approximately 9.5 percentage points. These results underscore the value of reconstructing temporally masked low-confidence trajectories for learning robust autism behavior representations. Taken together, the findings suggest that the proposed confidence-aware temporal masking strategy enables the Transformer encoder to capture richer behavioral dynamics, leading to improved representation quality and downstream classification performance.

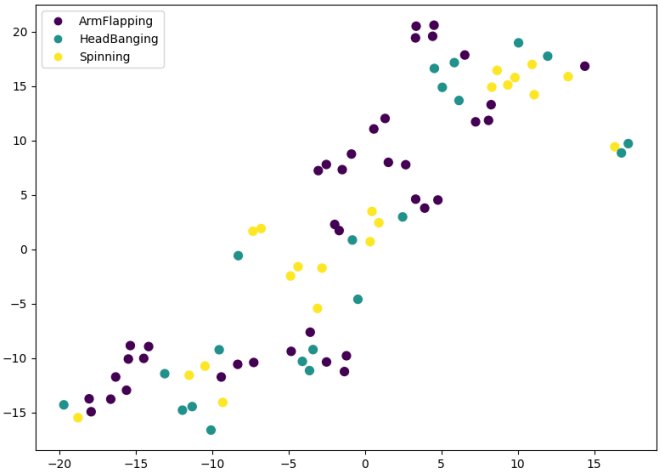


Fig. 5. Visualization of behavior representations learned by the pretrained Temporal CAMML encoder.

V. CONCLUSION

In this paper, we propose Temporal CAMML, a confidence-aware temporal motion masking framework for self-supervised autism behavior representation learning from skeletal pose sequences. This work presents a self-supervised framework for learning motion representations from autism behavior videos without relying on behavior annotations during pretraining. The proposed Temporal CAMML approach combines confidence-guided joint selection, temporal trajectory masking, and Transformer-based reconstruction to capture meaningful behavioral dynamics from skeletal pose sequences. Experimental evaluation on demonstrated the effectiveness of the proposed strategy. The results suggest that leveraging pose confidence estimates together with temporal motion information can improve the analysis of autism behaviors, particularly in settings where labeled data are limited. Future research will investigate the scalability of the approach on larger autism behavior datasets and explore more expressive temporal modeling techniques, including graph-based architectures. Another promising direction is the integration of multiple behavioral modalities, such as skeletal motion, facial expressions, and contextual scene information, to support richer representation learning and a more comprehensive understanding of behavioral patterns.

AI Use Disclosure

LLMs were used only to improve the language, clarity, and readability of the manuscript. The authors take full responsibility for the content.

References

[1] M. Dia, G. Khodabandelou, A. Q. M. Sabri, and A. Othmani, "Video-based continuous affect recognition of children with Autism Spectrum Disorder using deep learning," *Biomedical Signal Processing and Control*, vol. 89, Art. no. 105712, 2024.

[2] A. Jeyarani R. and R. Senthilkumar, "DSVTN-ASD: Detection of stereotypical behaviors in individuals with Autism Spectrum Disorder using a Dual Self-Supervised Video Transformer Network," *Neurocomputing*, vol. 624, Art. no. 129397, 2025.

[3] R. R. Jha, A. Muralie, M. Daroch, A. Bhavsar, and A. Nigam, "Enhancing Autism Spectrum Disorder identification in multi-site MRI imaging: A multi-head cross-attention and multi-context approach for addressing variability in un-harmonized data," *Artificial Intelligence in Medicine*, vol. 157, Art. no. 102998, 2024.

[4] A. Abu-Doleh, I. F. Abu-Qasmieh, H. H. Al-Quran, I. S. Masad, L. R. Banyissa, and M. A. Ahmad, "Recognition of autism in subcortical brain volumetric images using autoencoding-based region selection method and Siamese Convolutional Neural Network," *International Journal of Medical Informatics*, vol. 194, Art. no. 105707, 2025.

[5] A. Arunkumar and S. Sudhakar Ilango, "DenWide-Net: Dense WideResNet for autism spectrum disorder detection using MRI image," *Biomedical Signal Processing and Control*, vol. 113, Art. no. 108976, 2026.

[6] X. Wang, Y. Fang, Q. Wang, P.-T. Yap, H. Zhu, and M. Liu, "Self-supervised graph contrastive learning with diffusion augmentation for functional MRI analysis and brain disorder detection," *Medical Image Analysis*, vol. 101, Art. no. 103403, 2025.

[7] S. Roy, R. Jadhav, and T. Meena, "Contrastive learning strategies for better image classification with imbalanced datasets," *Applied Soft Computing*, vol. 190, Art. no. 114450, 2026.

[8] L.-J. Marhuenda, F. Gomez-Donoso, and M. Cazorla, "HARTO: Human Activity Recognition Through Optical Flow Architecture," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2024.

[9] J. Lopez-Martinez, P. Checa, J. M. Soto-Hidalgo, I. Triguero, and A. Fernandez, "A Wearable Eye-Tracking Approach for Early Autism Detection with Machine Learning: Unravelling Challenges and Opportunities," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2024.

[10] M. Zeraati and A. Davoodi, "ASD-GraphNet: A novel graph learning approach for Autism Spectrum Disorder diagnosis using fMRI data," *Computers in Biology and Medicine*, vol. 196, Art. no. 110723, 2025.

[11] N. J. Dhinagar, V. Santhalingam, K. E. Lawrence, E. Laltoo, and P. M. Thompson, "Few-Shot Classification of Autism Spectrum Disorder using Site-Agnostic Meta-Learning and Brain MRI," in *Proc. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024.

[12] Anonymous Authors, "Balanced Learning with Token Selection for Few-Shot Classification," *Under Review at ICLR*, 2024.

[13] S. S. Rajagopalan, A. Dhall, and R. Goecke, "Self-Stimulatory Behaviours in the Wild for Autism Diagnosis," in *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*, Sydney, Australia, 2013, pp. 755–761, doi: 10.1109/ICCVW.2013.103.

[14] S. Donné, J. Davis, F. Van Utterbeeck, and M. Verbeke, "Anomaly Detection Under Contaminated Data: A Weighted Iterative Refinement Framework for Health Monitoring," in *Proc. IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, Birmingham, United Kingdom, Oct. 2025, doi: 10.1109/DSAA65442.2025.11248011.

[15] C. Smailis, C. Ordonez, and I. A. Kakadiaris, "Towards LLM-Guided Healthcare Dataset Harmonization," in *Proc. IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, Birmingham, United Kingdom, Oct. 2025, doi: 10.1109/DSAA65442.2025.11247965.

[16] T. Akalya, D. Ramyachitra, M. Shabarna, and C. Legha, "Deep Learning-Based Detection of Autism Spectrum Disorder and Emotion Recognition in Children," *Pattern Recognition*, vol. 173, Art. no. 112906, 2026.

[17] V. Chawla and B. Rana, "Early Fusion for Autism Classification with Biomarker Detection Using rs-fMRI and Phenotypic Data," *Biomedical Signal Processing and Control*, vol. 109, Art. no. 108020, 2025.

[18] L. Xiao, S. Zhang, S. Wu, H. Huang, and Z. Hu, "Classification of Autism Spectrum Disorders Based on the Adaptive Fuzzy Reasoning System," *Biomedical Signal Processing and Control*, vol. 104, Art. no. 107513, 2025.

[19] R. Parvathy, R. Arunachalam, S. Damodaran, M. Al-Razgan, Y. A. Ali, and Y. J., "An Intellectual Autism Spectrum Disorder Classification Framework in Healthcare Industry Using ViT-Based Adaptive Deep Learning Model," *Biomedical Signal Processing and Control*, vol. 106, Art. no. 107737, 2025.

[20] M. Dileo, M. Zignani, and S. Gaito, "A Discrete-Time Deep Learning Framework for Temporal Heterogeneous Networks Forecasting," in *Proc. IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, Birmingham, United Kingdom, Oct. 2025, doi: 10.1109/DSAA65442.2025.11416805.

Appendix A
Additional Pose Extraction Examples

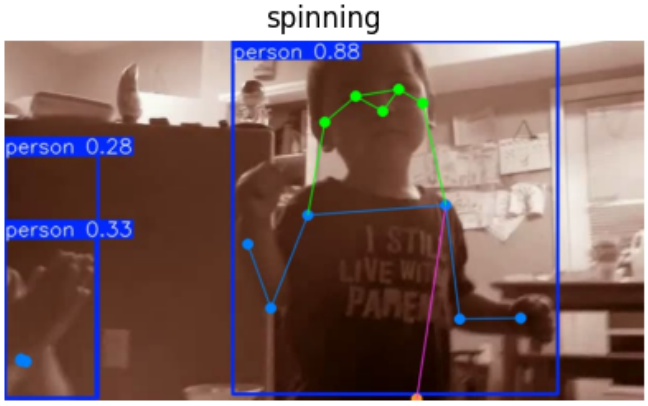


Fig. 6. Additional examples of pose extraction from SSBD behavioral events.

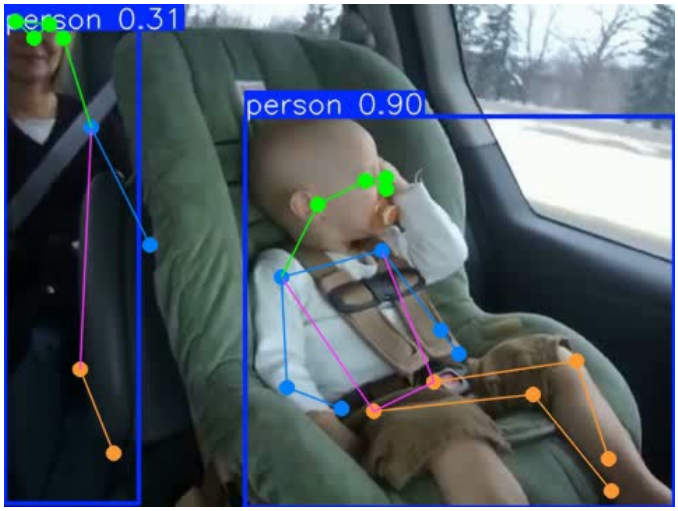


Fig. 7. Additional examples of pose extraction from SSBD behavioral events.